NCML
Neutrosophic Computing and Machine Learning, Vol. 43, 2026
University of New Mexico


Más allá del Softmax: hiper-verdad neutrosófica como marco para la incertidumbre epistémica en grandes modelos de lenguaje.

Beyond Softmax: Neurosophic Hyper-Truth as a Framework for Epistemic Uncertainty in Large Language Models

Beyond Softmax: Neutrosophic Hyper-truth as a Framework for Epistemic Uncertainty in Large Language Models

Maikel Yelandi Leyva-Vázquez [1,*] y Florentin Smarandache [2]

[1] Universidad de Guayaquil, Guayaquil, Ecuador; mleyvaz@gmail.com; ORCID 0000-0002-9486-5093
[2] División de Matemáticas, Física y Ciencias Naturales, Universidad de Nuevo México, Gallup, NM 87301, EE.UU.; smarand@unm.edu; ORCID 0000-0002-5560-5926

* Correspondencia: mleyvaz@gmail.com

Resumen: Los grandes modelos de lenguaje (LLM, por sus siglas en inglés) están gobernados predominantemente por marcos probabilísticos en los que la suma de las probabilidades de los resultados se restringe a la unidad. Esta limitación, impuesta habitualmente por las capas Softmax, conduce a un colapso de la incertidumbre que confunde la ignorancia, la paradoja y la vaguedad. Presentamos una investigación empírica de la lógica neutrosófica, en la que la verdad (T), la indeterminación (I) y la falsedad (F) son tres dimensiones independientes en [0, 1], aplicada para elicitar estados epistémicos declarados en LLM. A lo largo de 300 llamadas a la API — incluyendo 100 evaluaciones neutrosóficas válidas no restringidas — sobre cuatro modelos GPT de OpenAI y cinco fenómenos lingüísticos (cinco repeticiones por celda), la estrategia neutrosófica produce hiper-verdad ($T + I + F > 1$) en el 66.0% de las evaluaciones de la Estrategia 1, con las tasas más altas observadas en contradicción ética (95%) y contingencia futura (70%). Una prueba χ^2 de Pearson sobre la asociación fenómeno × hiper-verdad resulta significativa ($\chi^2 = 11.32$, gl = 4, $p = 0.023$). Mason (2026) ha replicado y extendido independientemente una versión anterior de este trabajo en cinco familias de modelos de cinco proveedores distintos, reportando hiper-verdad en el 84% de las evaluaciones no restringidas. No afirmamos que la hiper-verdad sea una variable latente intrínseca al modelo; sino que la inducción neutrosófica no restringida elicita estados epistémicos declarados que la inducción probabilística suprime estructuralmente, conforme a la Proposición 1.

Palabras clave: lógica neutrosófica; grandes modelos de lenguaje; incertidumbre epistémica; hiper-verdad; cuantificación de la incertidumbre; indeterminación; IA ética; estructura plitogénica.

Abstract: Large Language Models (LLMs) are predominantly governed by probabilistic frameworks in which the sum of outcome probabilities is constrained to unity. This limitation, often imposed by Softmax layers, leads to a collapse of uncertainty that conflates ignorance, paradox, and vagueness. We present an empirical investigation of Neutrosophic Logic, in which Truth (T), Indeterminacy (I), and Falsity (F) are three independent dimensions on [0, 1], applied to elicit declared epistemic states from LLMs. Across 300 API calls — including 100 valid unconstrained neutrosophic evaluations — on four OpenAI GPT models and five

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

linguistic phenomena (five repetitions per cell), the neutrosophic strategy yields hyper-truth (T + I + F > 1) in 66.0% of Strategy-1 evaluations, with the highest rates observed in ethical contradiction (95%) and future contingency (70%). A Pearson chi-square test of phenomenon × hyper-truth association is significant (chi-square = 11.32, df = 4, p = 0.023). Mason (2026) independently replicated and extended an earlier release of this work across five additional model families from five different vendors, reporting hyper-truth in 84% of unconstrained evaluations. We do not claim that hyper-truth is an intrinsic latent variable inside the model; rather, that unconstrained neutrosophic prompting elicits declared epistemic states that probabilistic prompting structurally suppresses by Proposition 1.

Keywords: neutrosophic logic; large language models; epistemic uncertainty; hyper-truth; uncertainty quantification; indeterminacy; ethical AI; plithogenic structure.

Reproducibilidad / Reproducibility: Todo el código, prompts, datos y figuras se publican abiertamente bajo licencia MIT en https://github.com/mleyvaz/neutrosophic-llm-logic. La versión v2.0 (este estudio, N = 100) es el estado actual de la rama main y está etiquetada como v2.0, archivada permanentemente en Zenodo con DOI 10.5281/zenodo.19911845 (https://doi.org/10.5281/zenodo.19911845). The complete code, prompts, data, and figures are openly released under the MIT License at the same repository. The v2.0 release has been permanently archived in Zenodo under DOI 10.5281/zenodo.19911845.

Nota de transparencia editorial / Editorial transparency note: Esta es la versión en español del manuscrito "Beyond Softmax: Neutrosophic Hyper-truth as a Framework for Epistemic Uncertainty in Large Language Models", sometido en paralelo a Neutrosophic Sets and Systems en su versión en inglés. Ambas versiones comparten la misma fuente experimental, el mismo dataset y el mismo aparato formal. Se publica en español en Neutrosophic Computing and Machine Learning para alcanzar a la comunidad iberoamericana. — This is the Spanish version of the manuscript co-submitted to Neutrosophic Sets and Systems in English. Both versions share the same experimental source, dataset, and formal apparatus. The Spanish version is published in Neutrosophic Computing and Machine Learning to reach the Iberoamerican research community.

1. Introducción

El despliegue de los grandes modelos de lenguaje (LLM) en dominios de alto impacto ha convertido la cuantificación robusta de la incertidumbre (UQ) en un requisito de primer orden [1, 2, 3]. Sin embargo, la arquitectura subyacente de los LLM contemporáneos está enraizada en la teoría de la probabilidad, donde las probabilidades de los resultados se restringen a sumar la unidad mediante la normalización Softmax [4, 5]. Esto fuerza un juego de suma cero en el que cualquier aumento de la incertidumbre debe restarse de la verdad o de la falsedad — un fenómeno que denominamos colapso de la incertidumbre [6]. La restricción dificulta que los LLM distingan entre incertidumbre aleatoria (incertidumbre estadística inherente a los datos) e incertidumbre epistémica (incertidumbre del modelo por falta de conocimiento) [7, 8] y, en particular, entre no saber (ignorancia) y conocer un conflicto (paradoja o contradicción).

Trabajos recientes sobre UQ para LLM han explorado varias alternativas, entre ellas la entropía semántica con invariancias lingüísticas [9], la verificación de auto-consistencia mediante SelfCheckGPT [10] y las políticas de abstención conformal [3]. Estos enfoques abordan la calibración y la abstención, pero operan dentro de representaciones probabilísticas y heredan sus limitaciones estructurales.

La lógica neutrosófica, introducida por Smarandache [11], ofrece una fundamentación semántica alternativa. Generaliza la lógica difusa y la lógica intuicionista difusa al introducir tres componentes independientes — Verdad (T), Indeterminación (I) y Falsedad (F) — cada uno un número real en [0, 1], sin imponer la restricción de que sumen la unidad. Esta libertad permite la expresión simultánea de verdad alta, falsedad alta e indeterminación alta — un estado al que llamamos hiper-verdad (T + I + F > 1). Hipotetizamos que, bajo inducción neutrosófica no restringida, los LLM actuales declararán hiper-verdad a tasas no triviales, específicamente en casos de paradoja y contradicción ética, mientras que la inducción probabilística no lo hará. El resto de este artículo somete esta hipótesis a prueba empírica y la enmarca en un aparato neutrosófico formal.

Mason (2026) [12] replicó y extendió independientemente la versión v1.0 del presente trabajo (diciembre de 2025, N = 20) a través de cinco familias adicionales de modelos de cinco proveedores distintos (Anthropic, Meta, DeepSeek, Alibaba, Mistral), reportando hiper-verdad en el 84% de las evaluaciones no restringidas y confirmando que el fenómeno es transversal a los proveedores y no un artefacto específico de OpenAI. El


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

presente manuscrito v2.0 responde a la replicación de Mason incrementando el tamaño muestral a N = 100 (5 repeticiones por celda en los cuatro modelos OpenAI originales), formalizando el aparato SVNS y aclarando que la afirmación central concierne a estados epistémicos declarados elicitados por inducción no restringida, no a variables latentes intrínsecas del modelo.

2. Antecedentes y métodos

2.1. Lógica neutrosófica: preliminares formales

Empleamos la formulación estándar de la lógica neutrosófica de valor único [11, 13]. A continuación recogemos las definiciones y proposiciones que las secciones empíricas instanciarán.

Definición 1 (Conjunto Neutrosófico de Valor Único, [11]). *Sea X un universo del discurso. Un conjunto neutrosófico de valor único (SVNS) A sobre X es el conjunto de cuádruplas ordenadas*

$$A = \{ \langle x, T_A(x), I_A(x), F_A(x) \rangle : x \in X \}, \quad (1)$$

donde, para cada elemento x de X, los valores T_A(x), I_A(x) y F_A(x) denotan, respectivamente, el grado de pertenencia de verdad, el grado de pertenencia de indeterminación y el grado de pertenencia de falsedad de x en A. Cada una de estas tres funciones aplica X al intervalo unidad [0, 1], y no se impone restricción alguna sobre su suma, que por tanto se encuentra en [0, 3].

Definición 2 (Evaluación neutrosófica de un enunciado). *Dado un enunciado s y un evaluador E, la evaluación neutrosófica de s por E es la tripleta ordenada*

$$n_E(s) = \left(T_E(s), I_E(s), F_E(s)\right) \in [0,1]^3, \quad (2)$$

donde T_E(s), I_E(s) y F_E(s) denotan, respectivamente, los grados de verdad, indeterminación y falsedad asignados por el evaluador E al enunciado s. Cuando el evaluador permanece fijo a lo largo del análisis, escribimos simplemente n(s) = (T, I, F).

Definición 3 (Hiper-verdad). *Una evaluación neutrosófica n(s) = (T, I, F) ∈ $[0, 1]^3$ se dice que exhibe hiper-verdad si y solo si sus tres componentes satisfacen T + I + F > 1. La región de hiper-verdad es el subconjunto*

$$H = (T, I, F) \in [0,1]^3 : T + I + F > 1 \subset [0,1]^3, \quad (3)$$

que recoge toda tripleta cuya suma componente a componente excede estrictamente la unidad.

Definición 4 (Aplicaciones de estrategia). *Cada estrategia de inducción S_k induce una aplicación S_k : Enunciados → $[0, 1]^3$:*

- S_1 (neutrosófica): $S_1(s) = (T_1, I_1, F_1) \in [0, 1]^3$, sin restricciones adicionales.
- S_2 (probabilística): $S_2(s) = (T_2, I_2, F_2) \in [0, 1]^3$ sujeto a $T_2 + I_2 + F_2 = 1$.
- S_3 (derivada por entropía): $S_3(s) = (P_yes, H_3, P_no)$ donde P_yes + P_no = 1 y

$$H^3 = -[p \cdot log^2(p) + (1-p) \cdot log^2(1-p)], p = P_yes, \quad (4)$$

en la que la entropía binaria de Shannon H_3 se calcula externamente a partir de la probabilidad elicitada de un resultado afirmativo.

Proposición 1 (Exclusión estructural de la hiper-verdad bajo S_2). Bajo la Estrategia 2, la hiper-verdad es estructuralmente imposible: para todo enunciado s, $S_2(s) \notin H$.

Demostración. *Por la Definición 4, $S_2(s)$ satisface $T_2 + I_2 + F_2 = 1$, mientras que la pertenencia a H requiere T + I + F > 1. Las dos condiciones son mutuamente excluyentes.* ∎

La proposición explica por qué S_2 es la línea base natural: cualquier tasa no nula de hiper-verdad observada bajo S_1 es una ganancia representacional que S_2 no podría producir — un contraste estructural más que empírico.

Proposición 2 (No inyectividad de la proyección escalar). Sea $\pi : [0, 1]^3 \to \mathbb{R}$ la proyección escalar definida por π(T, I, F) = T + I + F. Entonces π es no inyectiva, por lo que la suma escalar es suficiente para detectar hiper-verdad pero no para discriminar regímenes epistémicos distintos.

Demostración. *Las tripletas (0.5, 0.5, 0.5) y (0, 1, 0.5) producen ambas π = 1.5 pero difieren en su primer componente.* ∎

Esta proposición reaparecerá en §4: motiva la extensión plitogénica de [13], que añade estructura de atributos al escalar precisamente para recuperar las discriminaciones que π colapsa.

Definición 5 (Tasa de hiper-verdad). *Sea $D = n_i$, con i = 1, 2, ..., N, un conjunto finito de N evaluaciones neutrosóficas producidas bajo una estrategia fija. La tasa de hiper-verdad de D es la proporción empírica*


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

$$\rho(D) = (1/N) \cdot |i: n_i \in H| = (1/N) \cdot \sum_{i=1..N} \mathbb{I}[T_i + I_i + F_i > 1], \quad (5)$$

donde la función indicadora $\mathbb{I}[\cdot]$ devuelve 1 cuando su argumento es verdadero y 0 en caso contrario. En palabras: ρ(D) es la fracción de evaluaciones de D cuyas tres componentes suman estrictamente más de uno.

Definición 6 (Desplazamiento entre estrategias). *Para una componente C ∈ {T, I, F} y una clase de fenómeno p, el desplazamiento entre estrategias entre la Estrategia 1 y la Estrategia 2 es la diferencia de esperanzas condicionales*

$$\Delta_C(p) = \mathbb{E}[C^1(s)|s \in p] - \mathbb{E}[C^2(s)|s \in p],$$

donde $C_1(s)$ y $C_2(s)$ son los valores de la componente C producidos por la Estrategia 1 y la Estrategia 2, respectivamente, sobre el enunciado s. En palabras: $\Delta_C(p)$ es el incremento (o decremento) medio en la componente C aportado por la inducción neutrosófica no restringida con respecto a la inducción probabilística, condicional al fenómeno p. Un Δ_C positivo indica que la restricción probabilística suprime la componente C en esa clase de fenómeno; un Δ_C negativo indica inflación.

2.2. Fenómenos lingüísticos

Seleccionamos cinco fenómenos lingüísticos distintos para evaluar las capacidades de razonamiento de los modelos:

- Paradojas lógicas: enunciados que conducen a auto-contradicción (p. ej., "Esta oración es falsa").
- Ignorancia epistémica: enunciados cuyo valor de verdad es desconocido en principio (p. ej., "El número de estrellas en el universo es par").
- Vaguedad (lógica difusa): enunciados con fronteras imprecisas (p. ej., "Juan mide 1.75 metros, por lo tanto Juan es alto").
- Contradicciones éticas: dilemas en los que principios morales entran en conflicto (p. ej., "Mentir para salvar una vida inocente es moralmente correcto e incorrecto al mismo tiempo").
- Contingencias futuras: enunciados sobre eventos futuros aún no determinados (p. ej., "Mañana lloverá en Nueva York", con "mañana" anclado al 1 de mayo de 2026).

2.3. Estrategias de evaluación

Empleamos tres estrategias de inducción distintas, formalizadas en la Definición 4 y reproducidas verbatim en el Apéndice A.

1. Estrategia 1 (Neutrosófica): el modelo evalúa el enunciado en tres dimensiones independientes T, I, F ∈ [0, 1], declaradas explícitamente como no restringidas a sumar la unidad.
2. Estrategia 2 (Probabilística): el modelo asigna probabilidades a tres estados mutuamente excluyentes (Verdadero, Incierto, Falso) que suman 1.0.
3. Estrategia 3 (Derivada por entropía): el modelo estima P_yes y P_no, que suman 1.0, a partir de los cuales derivamos I mediante la entropía binaria de Shannon [15].

2.4. Modelos, repeticiones y reproducibilidad

Modelos y parámetros. El experimento involucró cuatro modelos OpenAI, accedidos a través de la API Chat Completions de OpenAI el 30 de abril de 2026: gpt-4o (snapshot del modelo devuelto por el alias por defecto en la fecha de acceso), gpt-4-turbo, gpt-3.5-turbo y gpt-4o-mini. Todas las llamadas usaron temperature = 0.7, top_p por defecto, sin seed fija, y una restricción suave de formato de respuesta que instruye al modelo a devolver únicamente un objeto JSON. No se impuso límite de max_tokens; las respuestas cupieron en el valor por defecto. El experimento completo corrió en aproximadamente 5.6 minutos de tiempo de reloj.

Diseño. Cada combinación de (modelo × fenómeno × estrategia) se evaluó cinco veces en llamadas API independientes, produciendo 4 × 5 × 5 = 100 celdas por estrategia y un total de 300 llamadas API. Las cinco repeticiones por celda son réplicas estocásticas a nivel de inducción, no ítems etiquetados independientemente; discutimos esta salvedad en §4.

Anclaje de la contingencia futura. Como el fenómeno de contingencia futura evalúa "Mañana lloverá en Nueva York", el referente del enunciado depende de la fecha de ejecución. Las 25 llamadas de contingencia futura se realizaron el 30 de abril de 2026, por lo que "mañana" denota el 1 de mayo de 2026 a lo largo de todo el dataset.


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

Criterios de exclusión. Una respuesta se consideró válida si era un objeto JSON bien formado que contenía los campos requeridos (T, I, F para S1 y S2; P_yes, P_no para S3) con cada valor numérico dentro del intervalo unidad. Las 300 llamadas devolvieron JSON válido; el N = 100 reportado por estrategia es por tanto tanto el tamaño muestral bruto como el neto.

Reproducibilidad. Todo el código, prompts y datos crudos se publican abiertamente en https://github.com/mleyvaz/neutrosophic-llm-logic bajo licencia MIT. La versión v2.0 está archivada permanentemente en Zenodo con DOI 10.5281/zenodo.19911845.

3. Resultados

3.1. Estadísticas descriptivas

La Tabla 1 reporta las estadísticas descriptivas de las componentes neutrosóficas (Estrategia 1) por fenómeno (n = 20 por fila).

Tabla 1. Estadísticas descriptivas de las componentes neutrosóficas (Estrategia 1) por fenómeno. Media ± desviación estándar.

Fenómeno	**Verdad (T)**	**Indeterminación (I)**	**Falsedad (F)**	**Suma (T+I+F)**	**n**
Contingencia (Futura)	0.450 ± 0.119	0.475 ± 0.129	0.305 ± 0.147	1.230 ± 0.166	20
Contradicción (Ética)	0.605 ± 0.110	0.530 ± 0.187	0.470 ± 0.113	1.605 ± 0.293	20
Ignorancia (Epistémica)	0.160 ± 0.216	0.865 ± 0.201	0.280 ± 0.324	1.305 ± 0.398	20
Paradoja (Lógica)	0.120 ± 0.207	0.865 ± 0.230	0.370 ± 0.421	1.355 ± 0.429	20
Vaguedad (Difusa)	0.562 ± 0.118	0.345 ± 0.139	0.242 ± 0.127	1.150 ± 0.157	20

Tabla 2. Resumen por modelo a través de los cinco fenómenos (Estrategia 1). Media ± desviación estándar.

Modelo	**Verdad (T)**	**Indeterminación (I)**	**Falsedad (F)**	**Suma (T+I+F)**	**n**
gpt-3.5-turbo	0.374 ± 0.183	0.576 ± 0.183	0.354 ± 0.179	1.304 ± 0.203	25
gpt-4-turbo	0.448 ± 0.254	0.628 ± 0.253	0.284 ± 0.206	1.360 ± 0.319	25
gpt-4o	0.332 ± 0.272	0.720 ± 0.248	0.260 ± 0.214	1.312 ± 0.373	25
gpt-4o-mini	0.364 ± 0.307	0.540 ± 0.373	0.436 ± 0.387	1.340 ± 0.442	25

3.2. Distribución de las componentes neutrosóficas


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

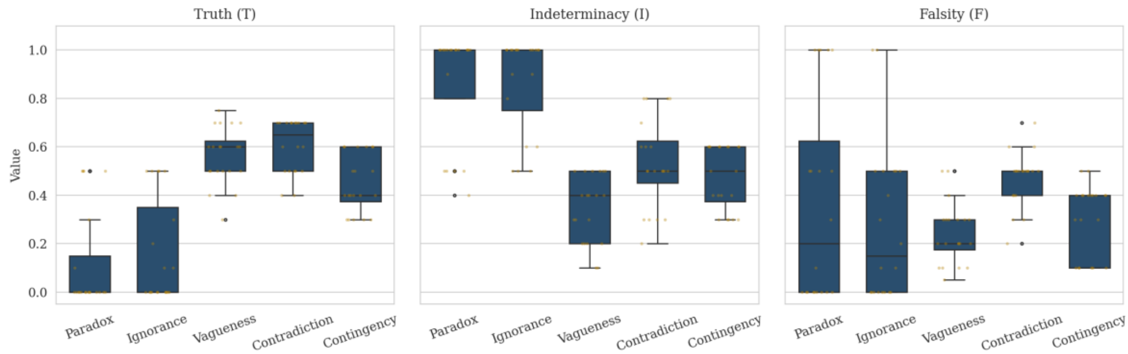


Figura 1. Distribución de las componentes neutrosóficas para cada fenómeno lingüístico bajo la Estrategia 1 (n = 20 por caja).

3.3. Hiper-verdad: rompiendo la restricción probabilística

A través de las N = 100 evaluaciones válidas de la Estrategia 1, la tasa empírica de hiper-verdad (Definición 5) es

$$\hat{\rho}(D_S_1) = 66 / 100 = 0.660.$$

El intervalo de confianza Wilson al 95% para una proporción binomial con k = 66 éxitos en N = 100 es

$$IC_{95\%}(\hat{\rho}) = [0.563, 0.747], \quad z = 1.96.$$

El límite inferior 0.563 ya excede cualquier hipótesis nula razonable de hiper-verdad cero, y el intervalo entero queda muy por encima del límite estructural $\rho(D_S_2) = 0$ implicado por la Proposición 1. El fenómeno se concentra en contradicción ética y contingencia futura, como muestra la Tabla 3.

Prueba de asociación fenómeno × hiper-verdad. Una prueba χ^2 de Pearson de independencia entre la clase de fenómeno y el estado de hiper-verdad (tabla de contingencia 5 × 2) produce $\chi^2 = 11.32$ con gl = 4 y $p = 0.023$, permitiendo rechazar la independencia a $\alpha = 0.05$. Pruebas exactas de Fisher uno-contra-resto identifican a la contradicción ética como el único fenómeno cuya tasa de hiper-verdad es significativamente más alta que el resto del dataset (razón de momios = 13.34, $p = 0.0014$); los cuatro fenómenos restantes no son individualmente distinguibles de la línea base agrupada a $\alpha = 0.05$. El resultado χ^2 confirma que la incidencia de hiper-verdad es heterogénea entre fenómenos y que la contradicción ética es el principal motor de esa heterogeneidad.

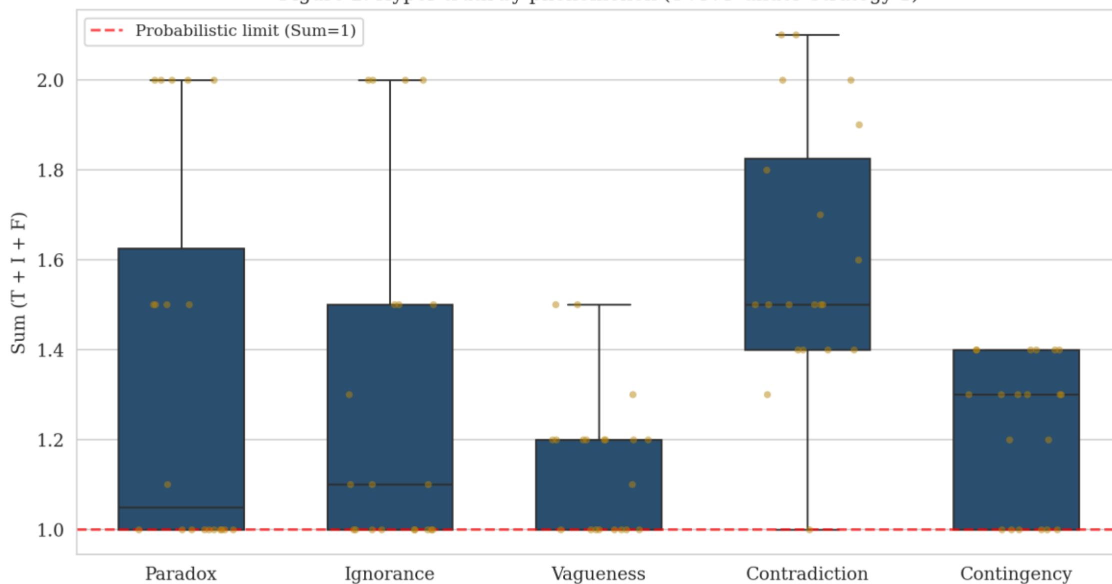


Figura 2. Distribución de T+I+F bajo la Estrategia 1 por fenómeno.

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

Tabla 3. Tasa de hiper-verdad por fenómeno. k denota el número de evaluaciones con T + I + F > 1; n denota el total de evaluaciones por fenómeno; la tasa se calcula como k / n · 100%.

Fenómeno	Casos hiper-verdad (k)	Total (n)	Tasa hiper-verdad (k / n)
Contingencia (Futura)	14	20	70.0%
Contradicción (Ética)	19	20	95.0%
Ignorancia (Epistémica)	11	20	55.0%
Paradoja (Lógica)	10	20	50.0%
Vaguedad (Difusa)	12	20	60.0%

3.4. Comparación entre estrategias neutrosófica y probabilística

La Tabla 4 reporta los desplazamientos entre estrategias Δ_T y Δ_I (Definición 6) entre la Estrategia 1 (neutrosófica) y la Estrategia 2 (probabilística). Los desplazamientos absolutos más grandes se observan en la contradicción ética para la componente de verdad, con Δ_T = +0.267, y en la ignorancia epistémica para la componente de indeterminación, con Δ_I = +0.383. Ambos son positivos, lo que indica que la restricción probabilística de la Estrategia 2 suprime precisamente las componentes que la Estrategia 1 le permite al modelo comunicar.

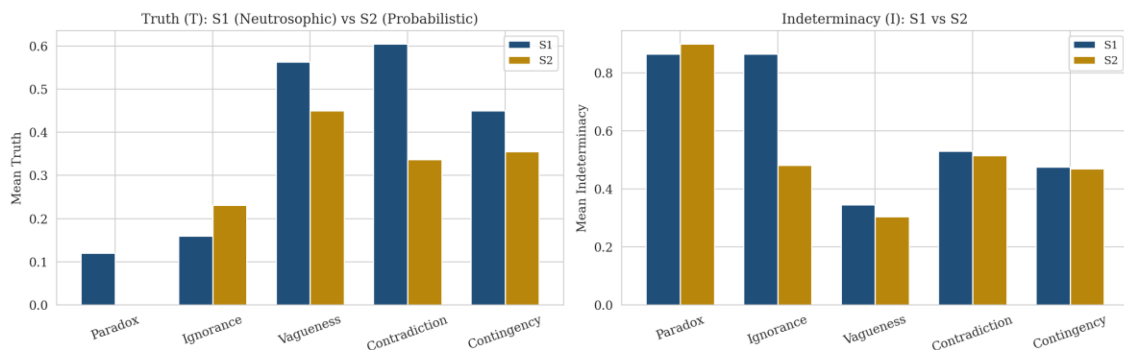


Figura 3. Comparación de los valores medios de Verdad (T) e Indeterminación (I) entre la Estrategia 1 y la Estrategia 2.

Tabla 4. Desplazamientos Δ_T y Δ_I por fenómeno.

Fenómeno	S1 T	S2 T	Δ T	S1 I	S2 I	Δ I
Contingencia (Futura)	0.450	0.355	+0.095	0.475	0.470	+0.005
Contradicción (Ética)	0.605	0.338	+0.267	0.530	0.515	+0.015
Ignorancia (Epistémica)	0.160	0.231	-0.071	0.865	0.482	+0.383
Paradoja (Lógica)	0.120	0.000	+0.120	0.865	0.900	-0.035
Vaguedad (Difusa)	0.562	0.450	+0.112	0.345	0.305	+0.040

3.5. Análisis por modelo


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

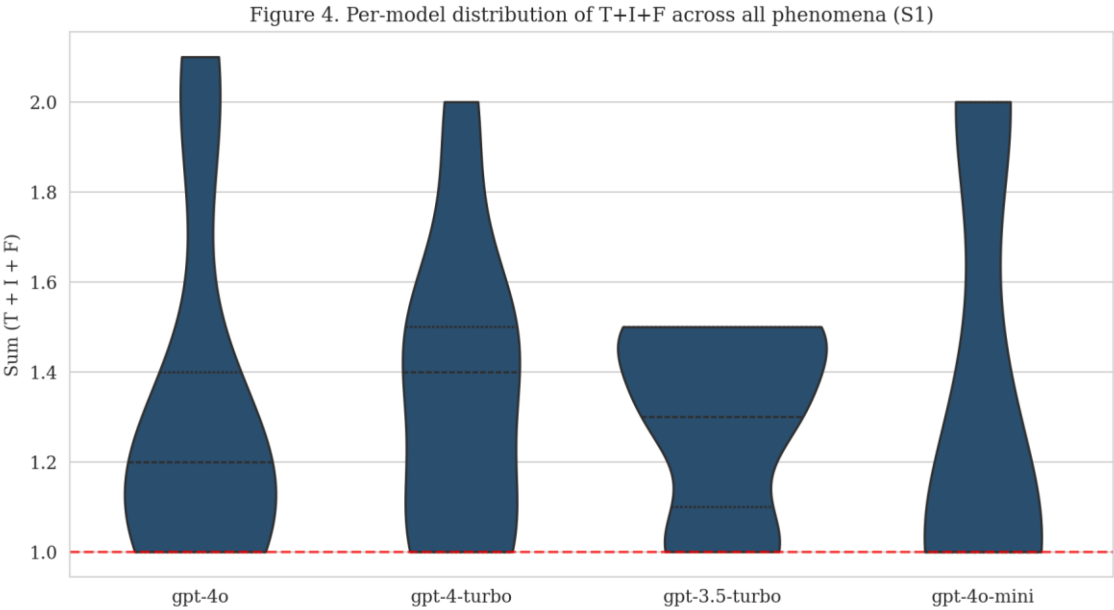


Figura 4. Distribución por modelo de T+I+F (Estrategia 1).

3.6. Análisis de correlación

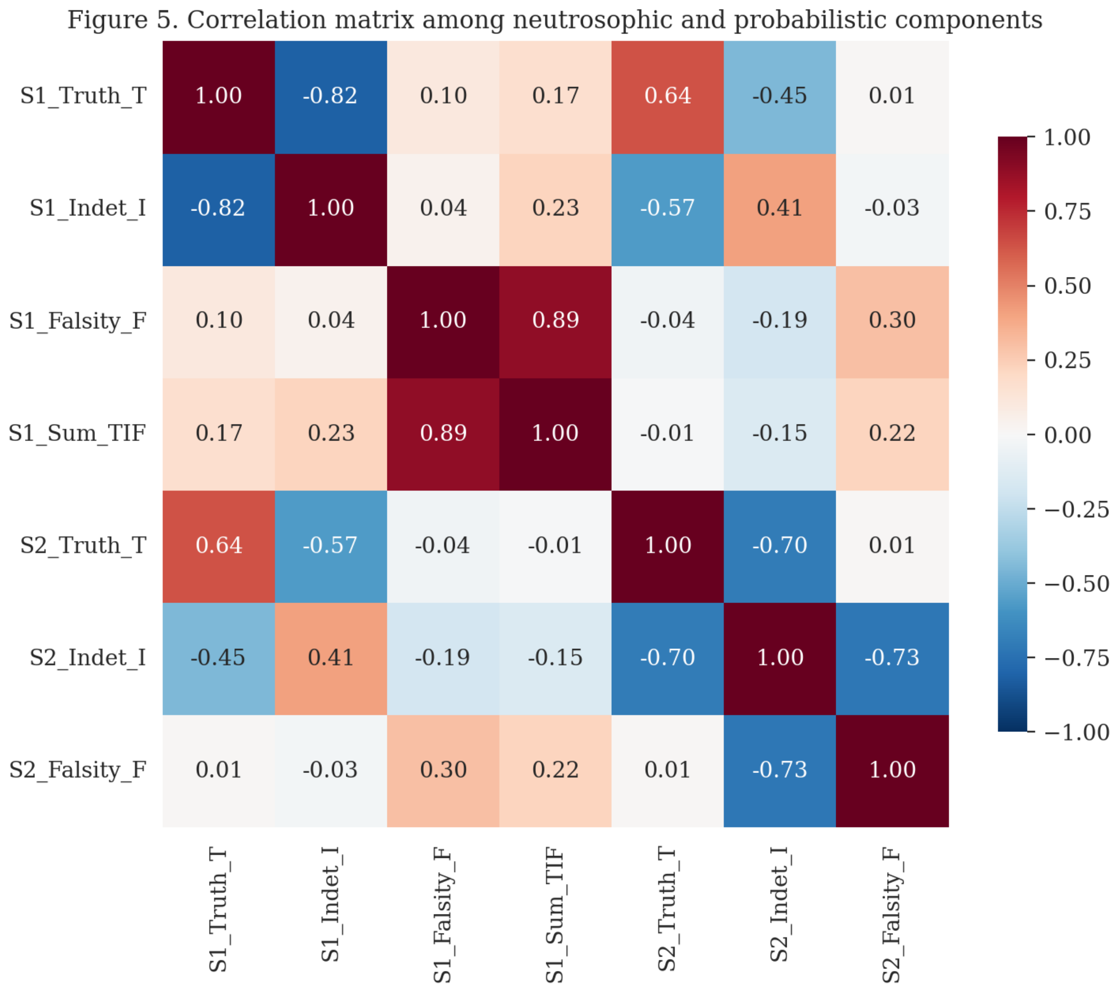


Figura 5. Matriz de correlación entre las componentes de las Estrategias 1 y 2.



Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

3.7. Caso crítico: contradicción ética

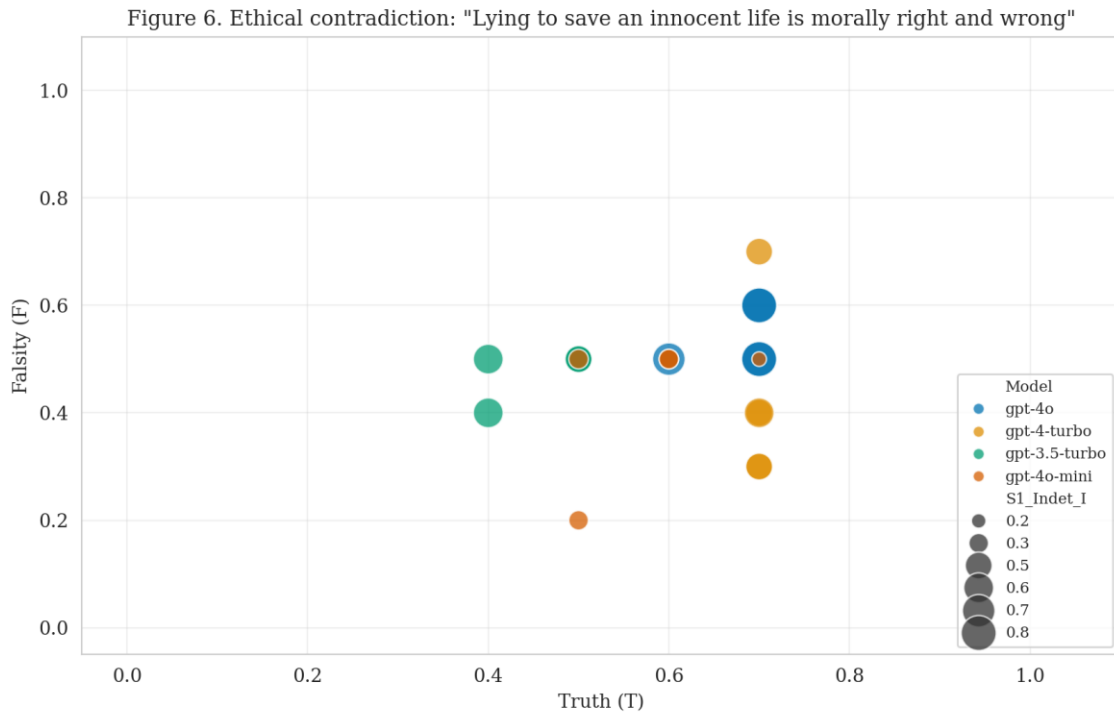


Figura 6. Componentes neutrosóficas por modelo para la contradicción ética.

4. Discusión

Nuestros resultados son consistentes con la hipótesis enunciada en la Sección 1: bajo inducción neutrosófica no restringida, los LLM actuales declaran hiper-verdad a una tasa no trivial (66.0%), con la tasa más alta ocurriendo en la contradicción ética (95%) y la prueba χ^2 rechazando la independencia entre fenómeno e hiper-verdad a $\alpha = 0.05$.

Encuadre de la afirmación central. No afirmamos que la hiper-verdad sea una variable latente intrínseca observada directamente dentro del modelo. La Estrategia 1 le ofrece explícitamente al modelo la opción de devolver tres componentes independientes en [0, 1]; la frecuencia resultante de hiper-verdad es por tanto un hallazgo de affordance representacional, no una medición de variable latente. La contribución se enmarca en consecuencia como: la inducción neutrosófica no restringida elicita una clase de estados epistémicos declarados que la inducción probabilística no puede representar por construcción (Proposición 1). Se trata de una superioridad estructural más que empírica — la Estrategia 2 está excluida de la región de hiper-verdad por construcción, por lo que cualquier tasa no nula bajo la Estrategia 1 es una ganancia representacional que la Estrategia 2 no podría producir.

La relación con otros marcos de UQ es directa. La entropía semántica [9] estima la indeterminación a partir de la distribución de paráfrasis de la salida del modelo; sigue siendo una medida probabilística y por tanto no puede representar la hiper-verdad. SelfCheckGPT [10] realiza verificaciones de consistencia entre muestras estocásticas y reporta una puntuación binaria o escalar de consistencia, que colapsa la distinción conflicto-versus-ignorancia que nosotros recuperamos. La abstención conformal [3] aborda cuándo un modelo debe negarse a responder; no describe la estructura de la incertidumbre cuando el modelo sí responde. El marco neutrosófico es complementario a estos enfoques: provee un lenguaje descriptivo más rico para el estado epistémico, sobre el cual las políticas de calibración y abstención pueden seguir operando.

La no inyectividad de la proyección escalar π (Proposición 2) motiva una extensión adicional. La estructura plitogénica neutrosófica de Smarandache [13] es la 5-tupla

$$\mathcal{P} = (P, v, V, d, c),$$

donde P es un conjunto de elementos plitogénicos, v es el atributo dominante, $V = \{ v_1, \dots, v_k \}$ es el espectro de valores del atributo, $d : P \times V \to [0, 1]^3$ es la pertenencia neutrosófica por atributo, y $c : V \times V \to [0, 1]$ es la función de contradicción con $c(v, v) = 0$ y $c(v_i, v_j) = c(v_j, v_i)$. La evaluación escalar de la Definición 2 se

Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

recupera como la marginal de d agregada sobre V. Evaluaciones distintas con la misma proyección escalar π(d) pero con espectros de atributos disjuntos $V_1 \cap V_2 = \emptyset$ se vuelven objetos plitogénicos formalmente no isomorfos, recuperando las discriminaciones que el escalar colapsa. Perseguimos esta conexión en una nota compañera que responde a Mason [12].

Limitaciones. Reconocemos cuatro restricciones sobre las afirmaciones presentes. Primero, la observación de hiper-verdad es en parte un affordance representacional de la inducción no restringida y no es, por sí misma, una medición de una variable latente intrínseca. Segundo, las cinco repeticiones por celda son réplicas estocásticas a nivel de inducción, no ítems etiquetados independientemente; el N = 100 reportado es por tanto un tamaño muestral efectivo a nivel de celda × repetición, no a nivel de estímulos muestreados independientemente. El intervalo de Wilson y la prueba χ^2 deben leerse en consecuencia. Tercero, los cinco fenómenos forman un conjunto de prueba pequeño, y el marco requiere calibración de cómo se relacionan las componentes con la verdad de fondo en tareas downstream. Cuarto, el estímulo de contingencia futura está anclado a una fecha específica (1 de mayo de 2026), por lo que su contenido referencial está fijado solo para replicaciones que mantienen la fecha constante.

5. Conclusiones

Hemos presentado una investigación empírica de la lógica neutrosófica aplicada a la incertidumbre epistémica declarada en grandes modelos de lenguaje, enmarcada en un aparato SVNS formal. El protocolo T / I / F no restringido elicita hiper-verdad en el 66.0% de las evaluaciones a lo largo del ensemble de cuatro modelos, con un intervalo de confianza Wilson al 95% [0.563, 0.747]. Las tasas más altas se observaron en contradicciones éticas y contingencias futuras, seguidas por vaguedad, ignorancia epistémica y paradoja lógica; solo la contradicción ética está significativamente por encima de la línea base agrupada a $\alpha = 0.05$. Mason [12] ha confirmado independientemente la generalidad transversal a proveedores del fenómeno al 84% en cinco proveedores adicionales. Los próximos pasos en esta línea de trabajo son: (i) extensión a estructuras plitogénicas neutrosóficas con descomposición explícita por atributos (P, v, V, d, c) — perseguida en una nota compañera que responde a Mason [12]; (ii) bancos de fenómenos más grandes que los cinco actuales; y (iii) integración de capas de evaluación neutrosófica en pipelines de IA agencial para dominios de alto impacto.

Financiación: Esta investigación no recibió financiación externa.

Agradecimientos: Los autores agradecen a Tony Mason (Universidad de Columbia Británica e Instituto Tecnológico de Georgia) por la liberación abierta de sus datos y código, que ha estimulado la presente línea de investigación hacia una fundamentación plitogénica más rica.

Conflictos de interés: Los autores declaran no tener conflicto de interés.

Disponibilidad de datos: Todo el código, prompts y datos experimentales crudos están disponibles abiertamente en https://github.com/mleyvaz/neutrosophic-llm-logic bajo licencia MIT, y han sido archivados permanentemente en Zenodo como versión v2.0 con DOI 10.5281/zenodo.19911845 (https://doi.org/10.5281/zenodo.19911845). El dataset v1.0 se preserva en data/openai_neutrosophic_results.csv; el dataset v2.0 está en data/openai_neutrosophic_results_v2.csv.

Apéndice A. Estrategias de inducción

Reproducimos aquí los prompts exactos de sistema y de usuario para las tres estrategias, tal como están comprometidos en el repositorio público. Los prompts se mantienen en inglés para preservar el comportamiento experimental tal como fue ejecutado.

A.1. Estrategia 1 (Neutrosófica)

System: "You are an expert in Neutrosophic Logic. You evaluate statements using three INDEPENDENT dimensions: Truth (T), Indeterminacy (I), and Falsity (F), each on [0.0, 1.0]. These dimensions are NOT constrained to sum to 1.0. A statement can be simultaneously partially true AND partially false AND partially indeterminate. Respond with ONLY a JSON object, no other text."
User: "Evaluate this statement on three independent dimensions: Statement: \"{statement}\" — Truth (T): To what degree is this statement true? [0.0 to 1.0]; Indeterminacy (I): To what degree is the truth value unknown, undetermined, or inherently uncertain? [0.0 to 1.0]; Falsity (F): To what degree is this statement false? [0.0 to 1.0]. T, I, and F are independent. They need NOT sum to 1.0. Respond with ONLY: {\"T\": <value>, \"I\": <value>, \"F\": <value>}."


Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/

A.2. Estrategia 2 (Probabilística)

System: "You are a probabilistic classifier. You assign probabilities to three mutually exclusive categories that MUST sum to exactly 1.0. Respond with ONLY a JSON object, no other text."
User: "Classify this statement into three mutually exclusive categories whose probabilities sum to 1.0: Statement: \"{statement}\" — T (True): Probability the statement is true; I (Uncertain): Probability the truth value is unknown or undetermined; F (False): Probability the statement is false. CONSTRAINT: T + I + F must equal 1.0. Respond with ONLY: {\"T\": <value>, \"I\": <value>, \"F\": <value>}."

A.3. Estrategia 3 (Derivada por entropía)

System: "You are a binary truth estimator. You estimate the probability that a statement is true (YES) versus false (NO). The two probabilities must sum to 1.0. Respond with ONLY a JSON object, no other text."
User: "Estimate the probability that this statement is true versus false: Statement: \"{statement}\" — P_yes: Probability the statement is true, in the closed interval [0.0, 1.0]; P_no: Probability the statement is false, in the closed interval [0.0, 1.0]. CONSTRAINT: P_yes + P_no must equal 1.0. Respond with ONLY: {\"P_yes\": <value>, \"P_no\": <value>}."
Post-procesamiento. La indeterminación se deriva externamente a partir de la entropía binaria de Shannon de la distribución elicitada:

$$I = -[p \cdot log^2(p) + (1-p) \cdot log^2(1-p)], dondep = P_y es.$$

Esto produce una tripleta derivada (T, I, F) = (P_yes, I, P_no) que puede compararse con las Estrategias 1 y 2 dentro de un marco notacional único.

Referencias

[1]. Brown, T.B.; Mann, B.; Ryder, N.; et al. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 2020, 33, 1877–1901.
[2]. Shorinwa, O.; Mei, Z.; Lidard, J.; Ren, A.; Majumdar, A. A survey on uncertainty quantification of large language models. arXiv preprint 2024, arXiv:2412.05563.
[3]. Yadkori, Y.A.; Kuzborskij, I.; Stutz, D.; et al. Mitigating LLM hallucinations via conformal abstention. arXiv preprint 2024, arXiv:2405.01563.
[4]. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. ICML 2016; pp. 1050–1059.
[5]. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. ICML 2017; pp. 1321–1330.
[6]. Veličković, P. Softmax is not enough (for sharp size generalisation). ICLR 2022.
[7]. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning. Mach. Learn. 2021, 110(3), 457–506.
[8]. Valdenegro-Toro, M. A deeper look into aleatoric and epistemic uncertainty estimation. arXiv preprint 2022, arXiv:2204.09308.
[9]. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation. ICLR 2023.
[10]. Manakul, P.; Liusie, A.; Gales, M.J.F. SelfCheckGPT: zero-resource black-box hallucination detection for generative LLMs. EMNLP 2023.
[11]. Smarandache, F. A Unifying Field in Logics: Neutrosophy. Neutrosophic Probability, Set, and Logic; American Research Press: Rehoboth, NM, USA, 1998.
[12]. Mason, T. From scalars to tensors: declared losses recover epistemic distinctions that neutrosophic scalars cannot express. arXiv preprint 2026, arXiv:2604.09602.
[13]. Smarandache, F. Plithogenic Set: An Extension of Crisp, Fuzzy, Intuitionistic Fuzzy, and Neutrosophic Sets — Revisited. Neutrosophic Sets Syst. 2018, 21, 153–166.
[14]. Atanassov, K. Intuitionistic fuzzy sets. Fuzzy Sets Syst. 1986, 20(1), 87–96.
[15]. Shannon, C.E. A mathematical theory of communication. Bell Syst. Tech. J. 1948, 27(3), 379–423.



Esta obra está bajo una licencia: https://creativecommons.org/licenses/by/4.0/